les cahiers de l'INRIA

Rendez-vous avec les sciences et technologies de l'information et de la communication

VISION COGNITIVE

Au-delà de l'apparence

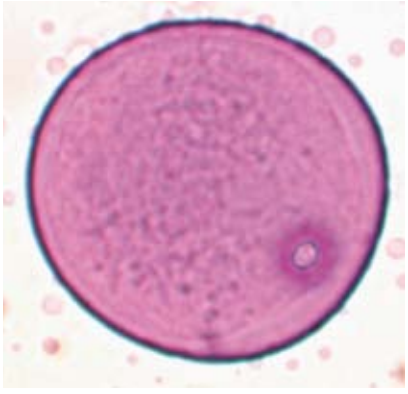
Il nous arrive de regarder sans voir, ou d'écouter sans entendre : pour interpréter les informations reçues par nos sens, il nous faut mettre en œuvre des processus cognitifs. Une machine peut-elle être dotée d'aptitudes similaires ?

Chacun sait qu'il existe une grande différence entre la perception d'un signal et la compréhension de son contenu. Dans le domaine sonore par exemple, nous pouvons entendre une personne parler sans comprendre un traître mot de ses paroles : si elle parle le chinois et que l'on ne maîtrise pas cette langue, les phrases prononcées seront audibles mais incompréhensibles. Dans le domaine visuel, cette différence entre le signal perçu (couleur, mouvement apparent, forme...) et ce qu'il représente est tout aussi importante. Quelles connaissances sur le monde physique doit-on intégrer dans un logiciel pour interpréter précisément des scènes réelles enregistrées par des caméras ? La vision cognitive s'intéresse précisément à ce décryptage des images[1].

La « lecture » d'une image suppose de disposer d'informations *a priori*, telles que les conditions de prises de vue, mais aussi de connaissances minimales relatives à l'environnement visualisé sur l'image. Sur le cliché ci-contre, par exemple, certains verront une simple représentation d'un objet peu identifiable, d'autres y reconnaîtront celle d'un grain de pollen allergisant de type *Poaceae* (fig. 1). Lorsque l'on cherche à automatiser la reconnaissance d'activités humaines ou d'événements dans le monde physique observés par des caméras vidéos, nous sommes confrontés au même type de problème : il faut passer du simple flux vidéo, autrement dit d'une suite d'images bidimensionnelles, à une interprétation dans le monde physique des comportements spatio-temporels des personnes ou des objets présents dans ces images.

Comment s'y prendre ? Des méthodes statistiques sont d'ores et déjà couramment utilisées pour identifier des individus, des animaux, des véhicules ou tout autre élément, notamment par comparaison avec des images déjà vues. Ces méthodes sont toutefois insuffisantes dans le cas de scènes complexes. Elles ne permettent pas de resituer correctement tous ces objets dans leur contexte spatial réel, à savoir tridimensionnel (3D), ni dans leur contexte temporel. Or la connaissance de leur dynamique est indispensable dès lors que l'on veut détecter certains types de comportements.

La communauté scientifique travaillant en vision par ordinateur est active dans le domaine de la détection du mouvement, du suivi d'objets mobiles et plus récemment dans l'analyse de trajectoires. Mais très souvent, ces analyses sont faites dans le plan image, autrement dit dans un espace bidimensionnel. Elles dépendent donc des conditions de prises de vues telles que la position ou l'orientation de la ou des caméra(s) ou de la largeur du champ de vue. Or la robustesse de la reconnaissance (par exemple celle d'activités humaines), c'est-à-dire la garantie que le système ne donnera pas une fausse alarme sur un éventuel comportement intéressant ou anormal, nécessite de calculer la dynamique des objets physiques dans l'espace spatio-temporel 4D (trois dimensions spatiales et une dimension temporelle). C'est l'une des difficultés de la vision cognitive.



PROJET EUROPÉEN ASTHMA

Fig. 1 : Image de grain de pollen de type Poaceae obtenue par microscopie optique.

Des applications très diverses

Une application concerne le prototype de réacteur de fusion nucléaire de Cadarache. Le projet Monitore (2008-2009), mené en partenariat avec le Commissariat à l'énergie atomique (CEA), vise en effet à identifier, sur la base d'images infrarouges, des phénomènes thermiques anormaux susceptibles de nuire à la stabilité du plasma[2]. Une autre application (2006), en partenariat avec l'Institut national de recherche agronomique (INRA), a consisté à étudier le comportement de guêpes de taille millimétrique (trichogrammes) utilisées comme agents de lutte biologique contre des insectes ravageurs[3] (pyrale du maïs...). L'INRIA a également participé au projet européen Caretaker (2006-2008) sur des applications à la vidéosurveillance dans les métros de Turin et de Rome[4].
F. B.

François Brémond est chargé de recherche dans l'équipe Pulsar, à l'INRIA Sophia Antipolis-Méditerranée.

Une seconde difficulté tient à la multiplicité d'interprétations possibles d'une même scène dynamique. Par exemple, les scènes montrées sur les figures 2 et 3 peuvent être interprétées différemment selon les connaissances *a priori* de l'observateur et en fonction de ses objectifs. Dans le premier cas (fig. 2), si l'observateur ne dispose d'aucune information particulière sur la localisation de la scène, il reconnaîtra une scène d'intérieur au sein de laquelle deux personnes se dirigent vers une porte..., guère plus. En revanche, un expert en vidéosurveillance capable d'identifier le lieu comme étant celui d'une agence bancaire, avec sa configuration spatiale et ses règles de sécurité spécifiques, interprétera la même scène comme une attaque de banque : il verra une personne non autorisée accéder avec un employé dans une zone interdite. De même, l'image de la figure 3 sera très parlante pour un gériatre mais pas pour un observateur naïf.

C'est pour appréhender les contenus sémantiques de telles images (leur signification réelle) qu'au sein de notre groupe de recherche*, nous proposons une approche couplant des techniques de vision par ordinateur, pour l'analyse spatio-temporelle des vidéos, et des techniques d'intelligence artificielle, pour la modélisation des connaissances *a priori*. Concrètement, cette approche se déroule en plusieurs étapes.

Supposons que nous nous intéressions à des comportements humains. La première étape est celle de la détection et du suivi des personnes présentes dans les scènes filmées par un nombre plus ou moins élevé de caméras. L'objectif est de les localiser et d'identifier leurs postures (fig. 4). Pour ce faire, nous utilisons deux catégories de connaissances a priori : une carte tridimensionnelle du lieu observé, d'une part, et des modèles tridimensionnels caractéristiques d'un certain nombre de postures, d'autre part (fig. 5).

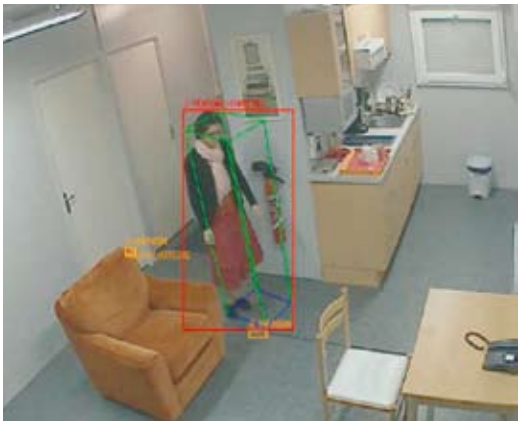
L'étape suivante consiste à mettre en correspondance, en temps réel, cette analyse spatio-temporelle 4D avec des modèles d'activités humaines prédéfinis. Une des difficultés réside bien sûr dans la formalisation de ces activités. Nous élaborons à ce stade des langages avec un vocabulaire, chaque élément de vocabulaire correspondant à une posture et à une localisation, et une grammaire, qui sont des compositions d'éléments du vocabulaire, autrement dit des activités. Le vocabulaire comprend un nombre limité de « mots », mais ils sont choisis de manière à être directement liés à la situation étudiée : ils sont déduits des mesures extraites de l'étape de détection et de suivi. Le langage ainsi construit permet d'exprimer des modèles d'activités élémentaires. Par exemple, il permettra d'exprimer le fait que la personne visualisée se trouve à proximité d'un fauteuil

© INRIA

Fig. 2 et 3 : à gauche, deux interprétations possibles : deux personnes marchant ensemble vers une porte, ou bien une attaque dans une banque. À droite, en l'absence d'information *a priori*, le profane voit une personne seule debout dans une pièce. L'expert médical reconnaît l'une de ses patientes âgées en train de préparer un repas dans sa cuisine.

* L'équipe projet Pulsar s'intéresse à l'interprétation sémantique et en temps réel de scènes dynamiques observées par des capteurs.

(fig. 4), ou bien qu'elle change de lieu et de posture...
Pour reconnaître des activités plus complexes, ces éléments doivent être combinés sous forme de séquences temporelles d'événements : par exemple, une personne entre puis se dirige vers le fauteuil, elle s'assoit et reste longtemps dans une position de repos, *etc.* Il est également possible d'exprimer des variantes dans le déroulement de l'activité grâce à des alternatives ou à des séquences optionnelles. Ces techniques ont d'autant plus d'intérêt que des raisonnements similaires peuvent être conduits pour tout objet physique, du monde vivant ou pas (voir l'encadré « Des applications très diverses »). Les résultats déjà obtenus intéressent le milieu industriel de la vidéosurveillance : une start-up, Keeneo a été créée en juillet 2005 pour l'industrialisation des techniques de notre équipe (voir l'encadré « Le bison futé des cantines »).
Un grand nombre de problèmes reste toutefois à résoudre. L'une des difficultés majeures concerne l'apprentissage incrémental (au fur et à mesure du traitement des informations) de modèles sémantiques d'activités. Sur quelles données d'apprentissage peut-on s'appuyer ? Quelles sont les parties des modèles connus à modifier ? Et surtout, comment garantir le lien entre l'information apprise par le logiciel et le monde physique ? Autant de questions qu'il va falloir résoudre à l'heure où l'interprétation sémantique d'activités humaines commence à intéresser un secteur aussi sensible que la santé. L'objectif est par exemple de pouvoir maintenir à domicile des personnes âgées et/ou malades (voir l'entretien avec Olivier Guérin), ce qui suppose de travailler en relation étroite avec les différents acteurs du domaine. M. T.



© INRIA

Fig. 4 et 5 : à gauche, la personne représentée est détectée en trois dimensions par son mouvement et son gabarit. L'image de droite correspond à l'interprétation de celle de gauche : la personne réelle est représentée par un modèle 3D de personne et replacée dans un modèle 3D du lieu observé par les caméras.

Monique Thonnat est directrice de recherche à l'Inria. Elle a créé successivement les équipes-projet Orion en 1995 et Pulsar en 2008 à Sophia Antipolis. Elle est aussi cofondatrice et conseiller scientifique de la société de vidéosurveillance Keeneo.

« Le bison futé des cantines »

Créée mi-2005, la start-up française Keeneo développe des logiciels d'analyse vidéo sur la base d'algorithmes mis au point par l'équipe de Monique Thonnat à l'Inria. Deux grands domaines d'application : l'analyse de flux, appliquée par exemple à la gestion de files d'attente, et la vidéosurveillance. Des systèmes de gestion de files d'attente sont installés dans des aéroports, mais aussi dans des lieux tels que la cantine du siège de la Société Générale (6 000 repas par jour) : l'idée est d'informer le personnel en temps réel sur leur temps d'attente. *« C'est ce que j'ai appelé le "bison futé" des cantines ! »*, note Thomas Herlin, directeur des ventes. En matière de sécurité, Keeneo propose des logiciels de détection d'intrusion pour des sites pétroliers, des distributeurs d'énergie, des constructeurs automobiles... L'entreprise a par ailleurs mis au point des systèmes de contrôle automatique de caméras, dont les applications concernent par exemple les milliers de caméras d'un réseau de métro urbain, et elle s'est engagée depuis peu dans des projets, notamment européens, de systèmes d'aide aux handicapés.

(1) M. Thonnat, *Semantic Activity Recognition, in* European Conference in Artificial Intellligence (ECAI), 2008.

(2) V. Martin *et al., Thermal Event Recognition Applied to Tokamak Protection during Plasma Operation, in* the IEEE International Instrumentation and Measurement Technology Conference, 2009.

(3) MB. Kaâniche *et al., Monitoring Trichogramma Activities from Videos : An Adaptation of Cognitive Vision System to Biology Field,* International Cognitive Vision Workshop, 2006.

(4) S. Hongeng *et al., Bayesan Framework for Video Surveillance Application,* Proc. of the 15th International Conference on Pattern Recognition, 2000.

D.R.

Entretien avec Olivier Guérin
Santé à domicile

La population française vieillit : une antienne. Les technologies de l'information et de la communication pourraient à la fois permettre d'améliorer le confort des personnes âgées et contribuer à réduire les frais d'hospitalisation.

Comment se posent les problèmes de maintien à domicile des personnes âgées ?
Olivier Guérin : En guise de préalable, il est bon de rappeler que le maintien à domicile est nettement moins coûteux pour la société que l'hospitalisation. Mais au-delà, notre objectif est surtout de préserver un maximum de bien-être à des personnes devenues vulnérables ou atteintes de pathologies liées au vieillissement. L'une des questions est alors de savoir comment les aider à rester autonomes tout en leur dispensant les soins nécessaires, voire en leur portant secours en cas d'accident. Cela pose le problème de leur surveillance en continu et de la mise en place d'un système d'alerte. Une difficulté tient au caractère très hétérogène de la population âgée. La méthode d'évaluation gériatrique standardisée, mise au point par le Californien Laurence Z. Rubinstein au milieu des années 1980, nous permet d'identifier les fragilités (nutritionnelle, cognitive, fonctionnelle...) de chaque individu et de mettre en place des stratégies d'intervention. Mais jusqu'à présent, ce type d'évaluation s'appuyait essentiellement sur des données subjectives (obtenues par exemple via des questionnaires). C'est l'une des raisons pour lesquelles nous nous intéressons aux technologies de l'information et de la communication : nous disposerons de données beaucoup plus objectives.

D'où le projet GerHome dont vous êtes partie prenante ?
O. G. : Ce projet, qui répondait à un appel à projets du conseil général des Alpes-Maritimes, visait plus généralement à concevoir et à expérimenter des solutions techniques pour le maintien à domicile. Entièrement financé par le conseil général et lancé en 2004, il associait en particulier le Centre hospitalo-universitaire (CHU) de Nice, l'Inria et le Centre scientifique et technique du bâtiment (CSTB). Il s'agissait en quelque sorte de réfléchir à un « domicile intelligent ». Le CSTB a eu en charge tout l'aspect relatif aux capteurs d'ambiance (capteurs de pression, d'ouverture de porte, d'intensité lumineuse...). L'Inria a apporté ses compétences en analyse d'images et en reconnaissance de comportements à risques dans des scènes réelles filmées. Nous avons commencé par construire un « appartement-laboratoire » équipé de divers capteurs et caméras et, fin 2007, a démarré la première phase d'étude clinique, c'est-à-dire de validation des technologies avec 14 volontaires sains (âgés de plus de 60 ans) mis en situation dans cet appartement pendant une journée. Nous sommes en train d'analyser les données recueillies. Il est clair que, dans ce contexte, l'analyse d'images nous apporte une aide irremplaçable.

L'étape suivante est donc de tester ces technologies avec des personnes malades ?
O. G. : Ce serait théoriquement la suite logique, sauf que placer des personnes vulnérables dans ce type de situation poserait des problèmes de sécurité sanitaire. Nous allons donc développer une plate-forme d'essai en milieu hospitalier (au sein du CHU de Nice). Cette possibilité nous est donnée grâce à la Direction générale des entreprises (DGE) qui nous a à son tour sollicités pour tester des technologies appliquées au maintien à domicile. Un nouveau projet a été mis en place, financé pour une part par la DGE et impliquant des partenaires publics et privés : le CHU de Nice et l'Inria, mais aussi l'université de Nice ainsi que le pôle de compétitivité mondial Solutions communicantes sécurisées (SCS). C'est sur cette plate-forme que nous allons démarrer dès cette année les tests sur des personnes fragiles ou malades.

Propos recueillis par Dominique Chouchan

Olivier Guérin exerce comme gériatre au sein de l'hôpital gériatrique Cimiez du Centre hospitalo-universitaire (CHU) de Nice. Il est également assesseur du doyen de la faculté de médecine de cette ville.

Réalisation rédactionnelle : Dominique Chouchan - Réalisation graphique : E-Graphics